BUILDING AN ANNOTATED CORPUS AND A LEXICAL DATABASE OF MODERN HEBREW IN XML

AUTHOR(S):

Sasaki, Tsuguya

CITATION:

Sasaki, Tsuguya. BUILDING AN ANNOTATED CORPUS AND A LEXICAL DATABASE OF MODERN HEBREW IN XML. 京都大学言語学研究 2004, 23: 17-45

ISSUE DATE:

2004-12-24

URL:

https://doi.org/10.14989/87846

京都大学言語学研究 (*Kyoto University Linguistic Research*) 23(2004), 17-45

BUILDING AN ANNOTATED CORPUS AND A LEXICAL DATABASE OF MODERN HEBREW IN XML*

Tsuguya Sasaki

1 Introduction

With the advent and ever-increasing implementation of Unicode, a 16-bit coded character set including most of the characters used in the major languages of the world, the application of computing is also making steady inroads on linguistics and Jewish studies, including the study of the Hebrew language. The interface between Hebrew linguistics and computer science, or Hebrew computational linguistics, can be approached from either of these two parent disciplines.[1] Taking the former as the starting point, the present paper proposes design principles for two electronic sources that may facilitate more empirical studies of the grammar and lexicon of Modern Hebrew, i.e., an annotated corpus and a lexical database, with XML (Extensible Markup Language)[2] as their storage (and interchange) format.

Section 2 below briefly surveys XML as well as its application to and/or potential for linguistic research, and explains why it is an ideal storage and interchange format for these purposes. Subsequently, sections 3 and 4 propose Hebrew-specific annotation schemes for a corpus and a lexical database[3] respectively. This paper does *not* deal with

* The present paper is based on the presentation I made in the session "Computing and Jewish Studies" at the 34th Annual Conference of the Association for Jewish Studies (December 2002, Los Angeles). I would like to express my gratitude to the following people (in alphabetical order) for their comments and suggestions on the handout of the presentation and the draft of the paper: Sarah Bunin Benor (Hebrew Union College), Shmuel Bolozky (University of Massachusetts Amherst), David Grossman (Michlalah Jerusalem College), Shlomo Izre'el (Tel Aviv University), Heidi Lerner (Stanford University), Ora Schwarzwald (Bar-Ilan University), Shuly Wintner (University of Haifa), Shlomo Yona (University of Haifa) and Ghil'ad Zuckermann (University of Haifa). Of course, I alone am responsible for all the mistakes that may remain.

1 Wintner (2003) is an excellent paper surveying the state of the art of Hebrew computational linguistics from the viewpoint of computer science.

2 See Bray et al. (2000); see also Sperberg-McQueen & Burnard (2002), Chapter 2 ("A Gentle Introduction to XML").

3 What is proposed as a lexical database here is only an organized repository of lexemes, hence should not be confused with a database management system.

the following issues though they are of course important: choice and size of primary linguistic sources on which they will be based; development and use of automation tools such as tokenizers, taggers, morphological analyzers and parsers; data retrieval; physical output for human consumption.

2 XML as a Document/Data Storage/Interchange Format

XML, often called "the ASCII of the 21st century", is a metalanguage for describing markup languages that are used in turn to mark up the logical structure of electronic texts. It is a simplified subset of SGML (Standard Generalized Markup Language), and preserves most of its power and richness, but retains all of its commonly used features and removes the more complex ones, thus making itself easier to extend. Its first version was issued in 1998, and then it was revised twice in 2000 and 2004. Unfortunately, however, its true value is not so widely recognized by researchers in the humanities, including linguistics and Jewish studies, though they can benefit enormously from its use in storing and interchanging data related to their research.

In spite of its far-reaching implications and potentials, the idea behind XML is very simple. Any character data is marked with a start tag and an end tag, which constitute an element, and secondary information, if necessary, is added as an attribute inside the start tag of that element. In the following example `noun` is an element, `number` is an attribute, and `singular` is its value; `<noun>` is a start tag, and `</noun>` is an end tag.[4]

```
<noun number="singular">language</noun>
```

Theoretically, any logical term can be used for elements and attributes, and this is why this markup language is called "extensible".

XML includes an ever-growing number of satellite technologies. XML Schema[5] and RELAX NG[6] are two major schema languages that define an XML vocabulary, then validate XML documents so that they may confirm to the same use of the same elements and attributes. XSLT (Extensible Stylesheet Language Transformations)[7] transforms XML documents into another XML vocabulary, and can be used to process and retrieve data from them. XSL (Extensible Stylesheet Language),[8] also known as XSL-FO

4 XML codes are written in a monospace font in this paper.
5 See Thompson et al. (2001) and Biron & Malhotra (2001).
6 See Clark & Murata (2001).
7 See Clark (1999).
8 See Adler et al. (2001).

(Extensible Stylesheet Language Formatting Objects), specifies formatting semantics or physical output of XML documents.

XML applications and documents are generally dichotomized into document-centric (also called narrative-centric or text-centric) and data-centric ones though the boundary between the two can be blurred. Document-centric documents, for which XML was first devised, inheriting the legacy of SGML, are not so well structured and are meant more for human consumption, while data-centric documents, for which database management systems have been used, are more rigidly structured and meant mainly for machine consumption. Corpora and lexical databases are examples of these two types respectively in the area of computational linguistics.

XML has at least the following five advantages as a document/data storage/interchange format for linguistic sources. First, it is machine- and human-readable as it uses text format and not binary format so that it can be read with any text editor. Secondly, it is crossplatform-compatible/portable as it is a non-proprietary public standard independent of any commercial factor and interest. Thirdly, it is crosslinguistically compatible/portable with Unicode as its default encoding; Unicode includes Hebrew letters and diacritics as well as the International Phonetic Alphabet and other special Latin characters generally employed in transcribing or transliterating Hebrew. Fourthly, it is self-descriptive in that texts are marked up structurally with semantic tags. Fifthly, multiple nesting is allowed so that it is easier to structure data in multiple layers. Of course, there are also disadvantages. One of them is that especially when a document is data-centric, many tags are used repetitively as the data is not structured in a tabular format, so this increases its size, but as computers have bigger and bigger memory size, hence have faster processing speed, this will not pause a serious problem. To insert repetitive tags can be a nuisance if done manually, but this is mostly automated by a growing number of editors and integrated development environments tailored for XML.

The application of XML to linguistic research has just started, hence there are not many vocabularies that use it as their format. TEI (Text Encoding Initiative)[9] and XCES (Corpus Encoding Standard for XML),[10] originally started as SGML applications, are two famous examples of linguistic and/or literary XML applications; they are annotation schemes for literary texts in general and corpora respectively. Among many annotated corpora of various languages, including International Corpus of English (written and

9 See Sperberg-McQueen & Burnard (2002).
10 See Ide & Suderman (2002).

spoken English around the world), Penn Treebank (written and spoken American English), Susan Corpus (written American English), Prague Dependency Treebank (Czech) and HPSG-based Syntactic Treebank of Bulgarian,[11] the last one is (still) exceptional in that it uses XML as its annotation format as of this writing.

3 Annotated Corpus: Document-Centric Application of XML

3.1 Existing and Planned Corpora of Modern Hebrew

To the best of my knowledge, there are four existing and planned corpora of Modern Hebrew: Bar-Ilan Corpus of Modern Hebrew headed by Prof. Yaacov Choueka, Corpus of Spoken Israeli Hebrew headed by Prof. Shlomo Izre'el,[12] Hebrew Corpora by Shlomo Yona and other researchers at the University of Haifa[13] and Treebank of Modern Hebrew headed by Prof. Eli Shamir.[14] The following table summarizes the main features of the four corpora:

Corpus	Sources	Notation	Annotation	Annotation Format	Phase	Availability
Bar-Ilan Corpus of Modern Hebrew	written language	orthography	unannotated	-	completed	publicly unavailable
Corpus of Spoken Israeli Hebrew	spoken language	phonological transcription with orthography	unannotated	-	in preparation	(will be) publicly available
Hebrew Corpora	written language	morpho-phonological transcription	lexical and morpho-syntactic	XML	experimental	publicly available
Treebank of Modern Hebrew	written language	morpho-phonological transcription	morpho-syntactic and syntactic	non-XML	in preparation	(will be) publicly available

Bar-Ilan Corpus of Modern Hebrew is the only existing corpus whose planned work has been completed, and consists of contemporary novels and newspaper articles of the 80's. Since it is stored in conventional Hebrew orthography, it is human readable, but

11 See <http://www.bultreebank.org>.
12 See Izre'el et al. (2001) and <http://www.tau.ac.il/humanities/semitic/cosih.html>.
13 See <http://cl.haifa.ac.il/~shlomo/corpora/>.
14 See Sima'an et al. (2001) and <http://www.cs.technion.ac.il/~winter/Corpus-Project/project-description.html>.

since it is unannotated, it is not always possible to retrieve those kinds of grammatical information that linguists need, even if they search it with מלי״ם, a sophisticated morphological algorithm for unvocalized Modern Hebrew texts developed by Prof. Choueka himself. It is unfortunate that this only existing corpus of Modern Hebrew has not been made accessible to the community of researchers.

Corpus of Spoken Israeli Hebrew, which is still in preparation, is both ambitious and innovative in that it is the first project to build a corpus of the spoken variety of Modern Hebrew, and it proposes rigorous statistical and analytical criteria for the representativeness of a corpus. It aims to collect 1,000 cells of 5,000 words per cell, totaling five million words. Considering such a huge size and lack of reliable tools for automatic annotation, it is understandable that the corpus is not planned to be annotated, at least in the initial stage of building.

Hebrew Corpora is an experimental project undertaken by a team of computer scientists. Although only a small collection of samples is available, it is probably the first attempt to annotate a corpus of Modern Hebrew; it includes lexical and morphosyntactic annotations. It is also innovative in that it uses XML as its format, though the actual scheme might require expansion and sophistication from a linguistic point of view.

Treebank of Modern Hebrew, which is undertaken by a team of leading Israeli computer scientists specializing in the natural language processing of Modern Hebrew, is the first project to build a treebank, or a syntactically annotated treebank, for Modern Hebrew. One of the most important contributions this project will surely make is the development of tools for automating morphosyntactic and syntactic annotations. It is therefore all the more unfortunate that it adopts with minor modifications the annotation scheme employed in the Penn Treebank,[15] which predates the advent of XML. It seems, therefore, that its morphosyntactic annotation is based too much on that for English devised by this English treebank. Its part-of-speech tag set as presented in Sima'an et al. (2001) might also require some refinement from a linguistic point of view to better reflect the structure of Modern Hebrew.

3.2 Features

The design of an annotated corpus proposed here is more from the viewpoint of a

15 See <http://www.cis.upenn.edu/~treebank/>.

linguist whose main interest in building it is not to build automation tools for computers but to use it for a more corpus-based description of the grammar of Modern Hebrew as it is. It might, therefore, seem rather naive to NLP-oriented computational linguists and turn out to be rather impractical to implement. Four levels of annotations are planned: 1) syntactic annotation, or parsing; 2) morphosyntactic annotation, or part-of-speech tagging; 3) lexical annotation, or lemmatization; 4) morphological annotation, or inflection tagging.

Of the seven maxims which should apply in the annotation of text corpora, as formulated by Leech (1993) and paraphrased by McEnery & Wilson (2001: 33-34), the first, sixth and seventh ones seem especially important: 1) it should be possible to remove the annotation from an annotated corpus and revert to the raw corpus; 6) annotation schemes should be based as far as possible on widely agreed and theory-neutral principles; 7) no annotation scheme has the a priori right to be considered as a standard.

The first maxim makes it preferable to store text in the original Hebrew orthography as encoded in Unicode (UTF-8) and not in Latin transcription, however easier it might be for computers to process. The sixth maxim is another reason to choose XML as the metalanguage for applications for corpus annotations. Unfortunately, XML is still a new standard, so there is only one scheme for corpus annotations that is publicly available, and it is only for morphosyntactic annotation - XCES. It was therefore necessary to formulate a custom-made annotation scheme specific to Modern Hebrew and define it with RELAX NG, which, incidentally, is not only easier to read and write but also more expressive than XML Schema. As the seventh maxim says, this is merely a proposal by an individual, hence must be reviewed by others and revised.

3.3 Structure[16]

3.3.1 Framework

The overall framework of a corpus scheme is as follows. After the XML declaration on the first line, `<?xml version="1.0" encoding="UTF-8"?>`, comes the root element `corpus`; when the corpus is split into multiple documents, each one of them will start with this element. Directly under it there are two elements `head` and `body`. Inside `head`

16 The latest version of the summary of this section, including its RELAX NG schema, is available at: <http://www.ts-cyberia.net/corpus_h.html>.

the meta information about the corpus or one of its parts is included, including the set of child elements `author`, `year`, `title`, `place` and `publisher` for books, the set `author`, `year`, `title`, `periodical` and `volume` for journal articles, or the set `author`, `date` and `periodical` for newspaper articles, and the `body` includes the actual texts. Its direct children are the repetitive elements `heading` and `para` [paragraph], which in turn includes the repetitive element `s` [sentence] with the attribute `page`.

```
<?xml version="1.0" encoding="UTF-8"?>
<corpus> <!-- root element -->
 <head>
  <author></author>
  <year></year>
  <title></title>
  <place></place>
  <publisher></publisher>
 </head>
 <body>
  <heading></heading> <!-- repetitive -->
  <para> <!-- repetitive -->
   <s page="">text</s> <!-- repetitive -->
  </para>
 </body>
</corpus>
```

Here is the summary of the elements and attribute enumerated so far:

Element	Explanation	Parent	Attributes	Children	Usage
corpus	corpus			head body	root element
head	head	corpus		author year title place publisher	
				author year title periodical volume	
				author date periodical	
author	author	head			
year	year	head			
date	date	head			

title	title	head			
place	place	head			
publisher	publisher	head			
periodical	periodical	head			
volume	volume	head			
body	body	corpus			
heading	heading	body			repetitive
para	paragraph	body			repetitive
s	sentence	para	page	*17	repetitive

Attribute	Explanation	Elements	Obligatory Values	Usage
page	page	s		

3.3.2 Syntactic Annotation (Parsing)

This level of annotation is the most problematic because of the very nature of XML: elements must always be linearly nested. In Modern Hebrew and in many other languages phrasal constituents like verb phrases can often be discontinuous, hence nonlinear. The compromise proposed here is between constituency and dependency at the phrasal level. Noun phrases, adjective phrases and adverb phrases are assigned the elements `np`, `adjp`, `advp` respectively, while verbal phrases are not interpreted in the conventional sense of the word; only the verbal core that can consist of one or two verbs with or without a conjunction but without a noun phrase it governs is redefined here as a verbal phrase and marked with `vp`. This scheme proposes to annotate syntactic argument structure with verbs as the core and other phrases as their satellites. For this reason prepositional phrases are not defined in a conventional manner, either; they are defined as units consisting of one or two prepositions or a preposition and a noun, and are marked with `prepp` as linking units between the verbal core and its satellites. At the clausal level, the elements `nc`, `adjc` and `advc` are proposed for noun, adjective and adverb clauses respectively.

When a noun phrase is an obligatory argument, it will be marked with the attribute `role` (syntactic role) with `subj` (subject) or `obj` (object) as its value, depending on its

17 See the last paragraph of the following section.

syntactic role in the sentence in question. No further distinction is made of objects. When a verb has some obligatory argument, its valency is indicated as an attribute with `valency`.

The following two tables summarize the elements and attributes proposed for syntactic annotation. Since it is impossible to determine a priori which constituents of word level can be children of clausal and phrasal levels, and both can be nested in other elements and inside themselves, the cells for parents and children are marked with an asterisk, indicating that they must be constrained only after working empirically with a sufficient amount of actual texts using this scheme.

Element	Explanation	Parents	Attributes	Children	Usage
nc	noun clause	*		*	
adjc	adjective clause	*		*	
advp	adverb clause	*		*	
np	noun phrase	*	role	*	
adjp	adjective phrase	*		*	
advp	adverb phrase	*		*	
vp	verb phrase	*		*	
prepp	prepositional phrase	*		*	

Attribute	Explanation	Elements	Obligatory Values	Usage
role	syntactic role	np	subj [subject] obj [object]	optional
valency	valency	v		

3.3.3 Morphosyntactic Annotation (Part-of-Speech Tagging)

One tends to think naively that the classification of words into parts of speech (or word classes) is a self-explanatory issue that was settled long ago, but this is far from the truth. As Evans (2000: 708) points out, modern practice has been to use distributional, i.e., morphological and syntactic, criteria in defining parts of speech, and these criteria vary from language to language. It seems, therefore, unfortunate that those working for the Treebank of Modern Hebrew "have tried to keep as close as possible to the English tag set used by the Penn tree-bank" (Sima'an et al. 2001: 353).

Evans (2000: 709-710) proposes the following four crosslinguistic guidelines in defining parts of speech: 1) define word classes on the basis of language-internal distributional criteria, both morphological and syntactic, noting problematic cases where morphological and syntactic criteria do not coincide; 2) map the prototype structure of these categories, identifying criteria for varying degrees of centrality, and assigning class members appropriately; 3) correlate, across languages, the classes so defined on the basis of the semantic and functional characteristics of their core members; 4) examine the distribution of matched classes cross-linguistically, and the degree of consistency with which words expressing particular types of meaning are assigned to a given class.

In accordance with these guidelines, the following parts of speech are provisionarily proposed here as separate XML elements for Modern Hebrew:[18] `n` (noun), `art` (article), `adj` (adjective), `card` (cardinal number), `ord` (ordinal number), `v` (verb), `adv` (adverb), `quant` (quantifier),[19] `pron` (pronoun), `proadj` (proadjective),[20] `proadv` (proadverb),[21] `prep` (preposition), `prep-art` (coalescence of preposition and article), `exist` (existential marker),[22] `q` (question marker),[23] `neg` (negative particle),[24] `comp` (complementizer),[25] `rel` (relativizer),[26] `conj` (conjunction), `interj` (interjection), and `punct` (punctuation).[27]

The following tables is a list of these elements together with the attributes they take, which will be explained in the following two sections. The cells of parents are marked with an asterisk as empirical study is required to decide which parent elements each one of the elements listed here can take.

18 For other sets of parts of speech for Modern Hebrew, see, e.g., Rosén (1977), Glinert (1989) and Schwarzwald (2001), who probably presents what is more or less considered a consensus among many linguists working on Modern Hebrew. This paper used her classification as a reference and expanded or fine-tuned it.

19 For example, הרבה and קצת, which precedes nouns.

20 For example, זה as in האיש הזה.

21 For example, פה and שם.

22 יש and אין.

23 האם and ־ן.

24 לא.

25 ש and כי.

26 ש and אשר.

27 Although punctituations are not a part of speech, this class is added here so that no character data inside a sentence may remain unmarked with an XML element.

Element	Explanation	Parents	Attributes
n	noun	*	lemma number gender state
art	article	*	bound
adj	adjective	*	lemma number gender state
card	cardinal number	*	gender state
ord	ordinal number	*	number gender
v	verb	vp	lemma person number gender tense mood valency
adv	adv	*	
quant	quant	*	
pron	pronoun	*	bound person number gender
proadj	proadjective	*	number gender
proadv	proadverb	*	
prep	preposition	*	bound
prep-art	preposition + article	*	bound
exist	existential marker	*	
q	question marker	*	
neg	negative particle	*	
comp	complementizer	*	bound
rel	relativizer	*	bound
conj	conjunction	*	
interj	interjection	*	
punct	punctuation	*	

3.3.4 Lexical Annotation (Lemmatization)

Of the above parts of speech, open-class content lexemes, i.e., nouns, adjectives and verbs have the attribute `lemma`, whose value will be the so-called citation form of each nominal, adjectival or verbal lexeme. Although the character data in the other parts of the corpus will be unvocalized, lemmata will be rendered in an auxiliarily vocalized full spelling (כתיב מלא עם ניקוד מסייע).[28]

Attribute	Explanation	Elements
lemma	lemma	n adj v

3.3.5 Morphological Annotation (Inflection Tagging)

When a lexeme is a bound form, it will have the attribute `bound`, whose value is either `pre` (prefixial) or `suf` (suffixal). For nominal, adjectival and verbal lexemes the following pieces of inflectional information will be added as attributes: `person`, `number`, `gender`, `state`, `tense` and `mood`.

Attribute	Explanation	Elements	Obligatory Values	Usage
bound	boundness	art pron prep prep-art comp rel	pre [prefixal] suf [suffixal]	optional
person		v pron	1 2 3	optional

28 That is, consonantal skeletons remain the same as in the unvocalized full spelling; vocalization and diacritic signs are added minimally to distinguish phonemes; דָגֵשׁ is omitted except in distinguishing ב, כ and פ from ב, כ, and פ respectively; placed on the upper right of שִׁין is omitted as it is less marked in frequency than שִׂין; and שׁוּא is omitted when it has the phonological value of zero whether it is historically שׁוּא נָח or שׁוּא נָע.

number		n adj ord v pron proadj	sg [singular] pl [plural]	optional
gender		n adj card ord v pron proadj	m [masculine] f [feminine]	optional
state		n adj card	c [construct]	optional
tense		v	past pres [present] fut [future]	optional
mood		v	imp [imperative] inf [infinitive]	optional

3.4 Example

The following two sentences will be annotated as follows:

העברית נמנית על משפחת הלשונות השמיות. בידינו כל תעודות או ידיעות על ראשית התהוותה, אך ספר המקרא מעיד על כך שכבר בשנת 1200 לפני הספירה היתה זו שפה מגובשת ועשירה.

```
<?xml version="1.0" encoding="UTF-8"?>
<corpus>
 <head>
  <author>Bar-Navi, E. (ed.)</author>
  <year>1992</year>
  <title>האטלס ההיסטורי: תולדות עם ישראל מימי האבות עד ימינו</title>
  <place>Tel Aviv</place>
  <publisher>Yediot Aharonot</publisher>
 </head>
 <body>
  <heading page="200">הלשון העברית, שפה עתיקה-חדשה</heading>
  <para>
   <s page="200">
    <art bound="pre">ה</art>
    <n lemma="עברית" number="sg" gender="f">עברית</n>
    <v lemma="נמנה" number="sg" gender="f" tense="pres">נמנית</v>
    <prep>על</prep>
    <n lemma="משפחה" number="sg" gender="f" state="c">משפחת</n>
```

```
    <art bound="pre">ה</art>
    <n lemma="לשון" number="pl" gender="f">לשונות</n>
    <art bound="pre">ה</art>
    <adj lemma="שחי" number="pl" gender="f">שחיות</adj>
    <punct>.</punct>
   </s>
   <s page="200">
    <exist>אין</exist>
    <prep>ב</prep>
    <n lemma="יד" number="pl" gender="f">ידי</n>
    <pron number="pl" bound="suf">נו</pron>
    <quant>כל</quant>
    <n lemma="תעודה" number="pl" gender="f">תעודות</n>
    <conj>או</conj>
    <n lemma="ידיעה" number="pl" gender="f">ידיעות</n>
    <prep>על</prep>
    <n lemma="ראשית" number="sg" gender="f" state="c">ראשית</n>
    <n lemma="התפתחות" number="sg" gender="f" state="c">התפתחות</n>
    <pron person="3" number="sg" gender="f" bound="suf">ה</pron>
    <punct>,</punct>
    <conj>אך</conj>
    <n lemma="ספר" number="sg" gender="m" state="c">ספר</n>
    <art bound="pre">ה</art>
    <n lemma="מקרא" number="sg" gender="m">מקרא</n>
    <v lemma="העיד" number="sg" gender="m" tense="pres">מעיד</v>
    <prep>על</prep>
    <proadv>כך</proadv>
    <comp>ש</comp>
    <adv>כבר</adv>
    <prep bound="pre">ב</prep>
    <n lemma="שנה" number="sg" gender="f" state="c">שנת</n>
    <card>1200</card>
    <prep>לפני</prep>
    <art bound="pre">ה</art>
    <n lemma="ספירה" number="sg" gender="f">ספירת</n>
    <v lemma="היה" person="3" number="sg" gender="f" tense="past">היתה</v>
    <pron number="sg" gender="f">זו</pron>
    <n lemma="שפה" number="sg" gender="f">שפה</n>
    <adj lemma="מגובש" number="sg" gender="f">מגובשת</adj>
    <conj bound="pre">ו</conj>
    <adj lemma="עשיר" number="sg" gender="f">עשירה</adj>
    <punct>.</punct>
   </s>
  </para>
 </body>
</corpus>
```

4 Lexical Database: Data-Centric Application of XML

4.1 Existing Lexical Databases of Modern Hebrew

A lexical database is an organized inventory of the lexemes of a language and includes information about them at various structural levels such as orthography, phonology, morphophonology, morphology, morphosyntax, syntax and/or sematics. It is similar to an electronic dictionary, and the boundary is not always so clear; generally speaking, however, the former is more structured and includes more descriptive information that is often missing in the latter.[29]

There seems to be no lexical database for Modern Hebrew that is publicly available. Of all the existing electronic dictionaries of Modern Hebrew, the CD-ROM and online[30] versions of Choueka (1997) approximate a lexical database most closely, but there are of course many types of missing grammatical information required for linguists analyzing the grammatical and lexical structure of Modern Hebrew; one cannot search lexemes according to, e.g., roots, affixes, etc.

4.2 Features

Basing itself probably on Choueka (1997) as its major reference, the lexical database proposed here includes information about the phonology, inflection, morphosyntax (parts of speech), word-formation and syntactico-semantics (meaning) of Modern Hebrew open-class content lexemes, including nouns, adjectives, verbs and adverbs, vis-à-vis closed-class function lexemes.

4.3 Structure[31]

4.3.1 Framework

Unlike corpus documents, (each fragment document of) a lexical database is better structured. After the XML declaration on the first line comes the root element `lexicon`,

29 See Calzolari (1989).
30 See <http://www.ravmilim.co.il/>.
31 The latest version of the summary of this section, including its RELAX NG schema, is available at: <http://www.ts-cyberia.net/lexicon_h.html>.

and inside the root element is the repetitive element `entry`, which is a kind of wrapper for one lexeme with the following five child elements: `headword`, `partofspeech`, `inflection`, `wordformation` and `meaning`. Each module will be briefly explained in the following five sectitions.

```
<?xml version="1.0" encoding="UTF-8"?>
<lexicon>
 <entry> <!-- repetitive -->
  <headword></headword>
  <partofspeech></partofspeech>
  <inflection></inflection> <!-- for nouns, adjectives and verbs>
  <wordformation></wordformation>
  <meaning></meaning> <!-- repetitive -->
 </entry>
</lexicon>
```

Element	Parent	Obligatory Data	Children	Usage
lexicon			entry	root element
entry	lexicon		headword phonology partofspeech wordformation meaning	repetitive
headword	entry		unvocalized vocalized transcription accent	
partofspeech	entry	noun adjective verb adverb		
inflection	entry		gender construct [construct state] plural pluralconstruct [plural construct state] future present infinitive	

wordformation	entry		type root pattern base prefix suffix family	
meaning	entry		label definition translation valency example	

4.3.2 Headword

The element `headword` for lemmas includes four child elements: `unvocalized`, `vocalized`, `transcription` and `accent`. The following symbols will be used for transcribing the consonants and vowels of Modern Hebrew; there is a one-to-one correspondence between a phoneme and a symbol. The transcription of lexemes, in addition to their Hebrew orthographical forms in an auxiliarily vocalized full spelling (כתיב מלא עם ניקוד מסייע) and an unvocalized spelling, will be useful in searching, e.g., certain consonantal clusters.

```
<headword>
 <vocalized></vocalized>
 <unvocalized></unvocalized>
 <transcription></transcription>
 <accent></accent> <!-- for nouns, adjectives and adverbs; repetitive -->
</headword>
```

Element	Parent	Obligatory Data	Children	Usage
headword	entry		unvocalized vocalized transcription accent	
unvocalized	headword			
vocalized	headword			
transcription	headword			
accent	headword			

The following symbols are used to transcribe Modern Hebrew consonants and vowels:

	Bilabial	Labio-dental	Alveolar	Palato-alveolar	Palatal	Velar	Glottal
Nasal	m		n				
Plosive	p b		t d			k g	'
Affricate			c	č ǧ			
Fricative		f v	s z	š ž		x r	h
Approximant					j		
Lateral			l				

	Front	Central	Back
Close	i		u
Mid	e		o
Open		a	

Accent is a type of phonological information that is missing in the majority of the dictionaries of Modern Hebrew. Although Choueka (1997) is one of the few exceptions, the indication of accent is restricted to the singular forms of nouns. As Sasaki (1997) showed, there are the following eight types of combinations for nouns; although the majority of the native words belong to the first or, less frequently, third type, and this variation is observed mainly in words borrowed from other languages, the place of accent in the singular and plural is not predictable.

Singular	Plural	Examples
ultimate	ultimate	talmíd - talmidím
ultimate	penultimate	studént - studéntim
penultimate	ultimate	séfer - sfarím
penultimate	penultimate	dólar - dolárim
penultimate	antepenultimate	seméster - semésterim
antepenultimate	ultimate	univérsita - universita'ót
antepenultimate	penultimate	télefon - telefónim
antepenultimate	antepenultimate	ópera - óperot

The accent of adjectives is more predictable. All or most of the native adjectives have the ultimate accent in all the four forms (i.e., singular masculine, singular feminine, plural masculine and plural feminine), while those borrowed from other languages have mostly the penultimate accent in the singular, and the accent remains on the same syllable in the other three inflectional forms. It follows that the accent of nouns must be

indicated for the singular and plural, while for adjectives it is enough to indicate the accent in the first two of the four inflectional forms.

4.3.3 Morphosyntax

Since this lexical database will be restricted to four open-class parts of speech, the element `partofspeech` has one of the four as its obligatory data: `noun`, `adjective`, `verb`, or `adverb`.

Element	Parent	Obligatory Data
partofspeech	entry	noun adjective verb adverb

4.3.4 Inflection

The categories of gender, state, and number are necessary for nouns, hence the elements `gender`, `construct`, `plural` and `pluralconstruct`. As for adjectives recorded in the database in masculine singular, the feminine (singular) form is enough to predict its plural masculine and feminine forms, so only the element `feminine` is proposed. All the inflectional forms of verbs can generally be predicted automatically, but verbs in Pa'al have irregular cases; although the present forms are always predictable, there are a small number of cases where the present form is nonexistent, and this must also be indicated. Therefore, the elements `future`, `present` and `infinitive` are proposed for verbs.

```
<inflection>
 <gender></gender> <!-- for nouns; repetitive -->
 <construct> <!-- for nouns; optional, repetitive -->
  <unvocalized></unvocalized>
  <vocalized></vocalized>
  <transcription></transcription>
  <accent></accent>
 </construct>
 <plural> <!-- for nouns; optional, repetitive -->
  <unvocalized></unvocalized>
  <vocalized></vocalized>
  <transcription></transcription>
  <accent></accent>
 </plural>
 <pluralconstruct> <!-- for nouns; optional, repetitive -->
```

```
  <unvocalized></unvocalized>
  <vocalized></vocalized>
  <transcription></transcription>
  <accent></accent>
 </pluralconstruct>
 <feminine> <!-- for adjectives; repetitive -->
  <unvocalized></unvocalized>
  <vocalized></vocalized>
  <transcription></transcription>
  <accent></accent>
 </feminine>
 <future> <!-- for verbs; optional, repetitive -->
  <unvocalized></unvocalized>
  <vocalized></vocalized>
  <transcription></transcription>
 </future>
 <present> <!-- for verbs; optional, repetitive -->
  <unvocalized></unvocalized>
  <vocalized></vocalized>
  <transcription></transcription>
 </present>
 <infinitive> <!-- for verbs; optional, repetitive -->
  <unvocalized></unvocalized>
  <vocalized></vocalized>
  <transcription></transcription>
 </infinitive>
</inflection>
```

Element	Parents	Obligatory Data	Children	Usage
inflection	entry		gender construct plural pluralconstruct feminine future present infinitive	
gender	inflection	masculine feminine		for nouns; repetitive
construct	inflection		vocalized unvocalized transcription accent	for nouns; optional, repetitive
plural	inflection		vocalized unvocalized transcription accent	for nouns; optional, repetitive

pluralconstruct	inflection		vocalized unvocalized transcription accent	for nouns; optional, repetitive
feminine	inflection		vocalized unvocalized transcription accent	for adjectives; repetitive
future	inflection		vocalized unvocalized transcription	for verbs; optional, repetitive
present	inflection		vocalized unvocalized transcription	for verbs; optional, repetitive
infinitive	inflection		vocalized unvocalized transcription	for verbs; optional, repetitive
unvocalized	construct plural pluralconstruct feminine future present infinitive			
vocalized	construct plural pluralconstruct feminine future present infinitive			
transcription	construct plural pluralconstruct feminine future present infinitive			
accent	construct plural pluralconstruct feminine			

4.3.5 Word-formation

Word-formation is probably the most complicated area of Modern Hebrew grammar.

The first child element proposed here is `type`; the obligatory data for it can be one of the following seven types: 1) `primitive base`, a base which cannot be decomposed into smaller morphemes; 2) `root-pattern formation`, a non-linear word-formation involving two discontinuous morphemes, i.e., a root and a pattern; 3) `reduplication`, a morphological process in which the internal composition of a base is modified with the repetition of its certain segment; 4) `affixation`, addition of an affix (prefix or suffix) to a base; 5) `blending`, a process in which two bases coalesce into a single stem without any internal boundary; 6) `acronyming`, a non-linear process in which two or more bases are coalesced into a noun; 7) `conversion`, a process in which there is no formal change and only the part of speech alters.[32]

The next element `root`, which refers to a skeleton of consonants shared by all the bases formed from it excluding preformative and postformative consonants, has two child elements: `primary` and `secondary`, which refer to primary and seconary roots respectively. The former are those roots which do not presuppose the existence of other roots or nonverbal stems, while the latter are those roots derived either from primary roots or nonverbal stems through the reduplication of part of their radicals or from nonverbal stems through the expansion of a part or of all their consonants. Roots, whether primary or secondary, are rendered in two ways, hence the child elements `grapheme` and `morphophoneme`. The graphemic notation renders a root in Hebrew characters, while the morphophonemic notation renders it in a series of (morpho)phonemes: for example, ט-לפ-ן vs. t-lf-n, but י-ב-ש vs. j-B-š. Morphophonemes are in capital letters; B, for example is a morphophoneme with the alternation of b~v.[33] In both methods of notation roots are rendered as three-slot skeletons, and each slot, which can include up to three consonants, is called a radical. The element `secondary` has two additional child elements: `type` and `source`. The former has either `expansion` (of primary roots) or `extraction` (from nonverbal stems) as its obligatory data, and the latter indicates either a primary source or a nonverbal stem from which the secondary root in question is made.

The third child element of `wordformation` is `pattern`, one of the two discontinuous morphemes involved in root-pattern formation. Because of the large number of nominal and adjectival patterns, only verbal patters will be indicated here: `pa'al`, `nif'al`, `pi'el`, `pu'al`, `hitpa'el`, `hif'il` and `huf'al`.

The fourth element `base` is optional for the last five types of word-formation, and the

32 This is in accordance with Sasaki (2000: 11-43).
33 An exhaustive list of morphophonemic alternations in roots is found in Sasaki (2000: 49-60).

last two elements `prefix` and `suffix` are optional only for the fourth type of word-formation.

```
<wordformation>
 <type></type>
 <root> <!-- optional for root-pattern formation -->
  <primary> <!-- optional -->
   <grapheme></grapheme>
   <morphophoneme></morphophoneme>
  </primary>
  <secondary> <!-- optional -->
   <grapheme></grapheme>
   <morphophoneme></morphophoneme>
   <type></type>
   <source>
    <unvocalized></unvocalized>
    <vocalized></vocalized>
    <transcription></transcription>
   </source>
  </secondary>
 </root>
 <pattern></pattern> <!-- optional for verbs -->
 <base> <!-- optional, up to 2 -->
  <unvocalized></unvocalized>
  <vocalized></vocalized>
  <transcription></transcription>
 </base>
 <prefix> <!-- optional -->
  <unvocalized></unvocalized>
  <vocalized></vocalized>
  <transcription></transcription>
 </prefix>
 <suffix> <!-- optional -->
  <unvocalized></unvocalized>
  <vocalized></vocalized>
  <transcription></transcription>
 </suffix>
</wordformation>
```

Element	Parents	Obligatory Data	Children	Usage
wordformation	entry		type root pattern base prefix suffix	

type	wordformation	primitive base root-pattern formation reduplication affixation blending acronyming conversion		
root	wordformation		primary secondary	optional
pattern	wordformation	pa'al nif'al pi'el pu'al hitpa'el hif'il huf'al		for verbs
base	wordformation		vocalized unvocalized transcription	optional, repetitive
prefix	wordformation		vocalized unvocalized transcription	optional
suffix	wordformation		vocalized unvocalized transcription	optional
primary	root		grapheme morphophoneme	optional
secondary	root		grapheme morphophoneme type source	optional
grapheme	primary secondary			
morphophoneme	primary secondary			
type	secondary		expansion extraction	
source	secondary		vocalized unvocalized transcription	
vocalized	base prefix suffix source			

unvocalized	base prefix suffix source			
transcription	base prefix suffix source			

4.3.6 Syntactico-Semantics

The repetitive element `meaning` has five child elements: `label`, `definition`, `translation`, `valency` and `example`. The first element `label` refers to usage label, and has one of the four optional character data: `archaic`, `literary`, `colloquial` and `slang`. The next two elements, `definition` and `translation`, are for the explanation of the meaning in Hebrew and its translational equivalence in English. The element `valency` optionally indicates prepositions which verbs require for their obligatory objects. The last, repetitive element `example` is for giving examples of sentences that contain the lexeme in question.

```
<meaning> <!-- repetitive -->
 <label></label> <!-- optional -->
 <definition></definition>
 <translation></translation>
 <valency></valency> <!-- optional, repetitive -->
 <example></example> <!-- optional, repetitive -->
</meaning>
```

Element	Parents	Obligatory Data	Children	Usage
meaning	entry		label definition translation valency example	repetitive
label	meaning	archaic literary colloquial slang		optional
definition	meaning			
translation	meaning			
valency	meaning			optional, repetitive
example	meaning			optional, repetitive

4.4 Example

The following is a sample entry. Since some of the elements are irrelevant to this concrete nominal lexeme, those irrelevant elements are of course omitted here.

```xml
<?xml version="1.0" encoding="UTF-8"?>
<lexicon>
 <entry>
  <headword>
   <vocalized>פִרמוט</vocalized>
   <unvocalized>פרמוט</unvocalized>
   <transcription>firmut</transcription>
   <accent>ultimate</accent>
  </headword>
  <partofspeech>noun</partofspeech>
  <inflection>
   <gender>masculine</gender>
  </inflection>
  <wordformation>
   <type>root-pattern formation</type>
   <root>
    <secondary>
     <grapheme>פ-רמ-ט</grapheme>
     <morphophoneme>f-rm-t</morphophoneme>
     <type>extraction</type>
     <source>
      <vocalized>פוֹרמָט</vocalized>
      <unvocalized>פורמט</unvocalized>
      <transcription>format</transcription>
     </source>
    </secondary>
   </root>
  </wordformation>
  <meaning>
   <definition>הכנה של דיסקט או של דיסק קשיח לכתיבת נתונים עליו תוך
    חלוקתם באמצעות סימונים מגנטיים מתאימים</definition>
   <translation>formatting</translation>
  </meaning>
 </entry>
</lexicon>
```

5 Summary

In the present paper detailed schemes are proposed for an annotated corpus and a lexical database of Modern Hebrew. They are meant to be primary linguistic sources for more empirical studies of the grammatical and lexical structure of Modern Hebrew to shed light on aspects and phenomena hitherto unknown. XML (Extensible Markup Language) is chosen as their storage format because of its machine- and human-readability, crossplatform-compatibility, crosslinguistic-compatibility, self-descriptiveness and capability of nesting structure. The corpus will be annotated in four levels, i.e., syntactically, morphosyntactically, lexically and morphologically. The lexical database will include modules of morphosyntax, inflection, word-formation and syntactico-semantics. The data will be recorded in Unicode, whether in Hebrew characters or in Latin transcription. Although countless numbers of revisions have been made since the idea of building these two sources in XML was first born a few years ago, what is proposed here is essentially by one individual. It might, therefore, need minor (or even major) revisions and/or expansions.

References

Adler, S. et al. 2001. *Extensible Stylesheet Language (XSL) Version 1.0 - W3C Recommendation.* <http://www.w3.org/TR/xsl/>.

Biron, P. V. & Malhotra, A. 2001. *XML Schema Part 2: Datatypes - W3C Recommendation.* <http://www.w3.org/TR/xmlschema-2/>.

Bray, T. et al. (eds.). 2004[3]. *Extensible Markup Language (XML) 1.0 - W3C Recommendation.* <http://www.w3.org/TR/REC-xml>.

Calzolari, N. 1989. Computer-Aided Lexicography: Dictionaries and Word Data Bases. In: I. S. Bátori et al. (eds.). *Computational Linguistics: An International Handbook on Computer Oriented Language Research and Applications.* Berlin. 510-519.

Choueka, Y. (ed.). 1997. **רב-מילים תמילון השלם**. Tel Aviv.

Clark, J. (ed.). 1999. *XSL Transformations (XSLT) Version 1.0 - W3C Recommendation.* <http://www.w3.org/TR/xslt>.

Clark, J. & Murata, M. 2001. *RELAX NG Specification.* <http://www.relaxng.org/spec.html>.

Evans, N. 2000. Word Classes in the World's Languages. In: G. Booij et al. (eds.).

Morphology: An International Handbook of Inflection and Word-Formation 1. Berlin. 708-732.

Glinert, L. 1989. *The Grammar of Modern Hebrew*. Cambridge.

Ide, N. & Suderman, K. 2002. *XCES: Corpus Encoding Standard for XML*. <http://www.xml-ces.org/>.

Izre'el, S. et al. 2001. Designing CoSIH: The Corpus of Spoken Israeli Hebrew. *International Journal of Corpus Linguistics* 6: 171-197.

Leech, G. 1993. Corpus Annotation Schemes. *Literary and Linguistic Computing* 8: 275-281.

McEnery, T. & Wilson, A. 2001[2]. *Corpus Linguistics*. Edinburgh.

Rosén, H. B. 1977. *Contemporary Hebrew*. The Hague.

Sasaki, T. 1997. ההטעמה בעברית החדשה: סוגיה ומקורותיה. Paper read at Post-Congress on Problems of Teaching Hebrew. Jerusalem.

Sasaki, T. 2000. The Verb Formation of Modern Hebrew. PhD Dissertation, The Hebrew University of Jerusalem.

Schwarzwald, O. 2001. *Modern Hebrew*. Munich.

Sima'an, K. et al. 2001. Building a Tree-Bank of Modern Hebrew Text. *Traitement Automatique des Langues* 42: 347-380.

Sperberg-McQueen, C. M. & Burnard, L. (eds.). 2002. *TEI P4: Guidelines for Electronic Text Encoding and Interchange, XML Version*. <http://www.tei-c.org/P4X/>.

Thompson, H. S. et al. (eds.). 2001. *XML Schema Part 1: Structures - W3C Recommendation*. <http://www.w3.org/TR/xmlschema-1/>.

Wintner, S. 2003. בלשנות חישובית עברית: עבר ועתיד. In: S. Izre'el & M. Mendelson (eds.). **מדברים עברית: חקר הלשון המדוברת והשונות הלשונית בישראל**. Tel Aviv. 35-64.

BUILDING AN ANNOTATED CORPUS AND A LEXICAL DATABASE OF MODERN HEBREW IN XML

Tsuguya Sasaki

Abstract

In the present paper detailed schemes are proposed for an annotated corpus and a lexical database of Modern Hebrew. They are meant to be primary linguistic sources for more empirical studies of the grammatical and lexical structure of Modern Hebrew to shed light on aspects and phenomena hitherto unknown. XML (Extensible Markup Language) is chosen as their storage format because of its machine- and human-readability, crossplatform-compatibility, crosslinguistic-compatibility, self-descriptiveness and capability of nesting structure. The corpus will be annotated in four levels, i.e., syntactically, morphosyntactically, lexically and morphologically. The lexical database will include modules of morphosyntax, inflection, word-formation and syntactico-semantics. The data will be recorded in Unicode, whether in Hebrew characters or in Latin transcription. Although countless numbers of revisions have been made since the idea of building these two sources in XML was first born a few years ago, what is proposed here is essentially by one individual. It might, therefore, need minor (or even major) revisions and/or expansions.

(受理日 2004年6月18日　最終原稿受理日 2004年12月2日)